

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-53394

(43) 公開日 平成11年(1999) 2月26日

(51) Int.Cl.⁴

識別記号

F I

G 0 6 F 17/30

G 0 6 F 15/403

3 5 0 C

17/21

15/20

5 7 0 N

15/403

3 4 0 A

審査請求 未請求 請求項の数11 F D (全 11 頁)

(21) 出願番号 特願平9-219298

(22) 出願日 平成9年(1997) 7月29日

(71) 出願人 390024350

株式会社ジャストシステム

徳島県徳島市神浜東3-46

(72) 発明者 野村 直之

徳島県徳島市神浜東3丁目46番地 株式会

社ジャストシステム内

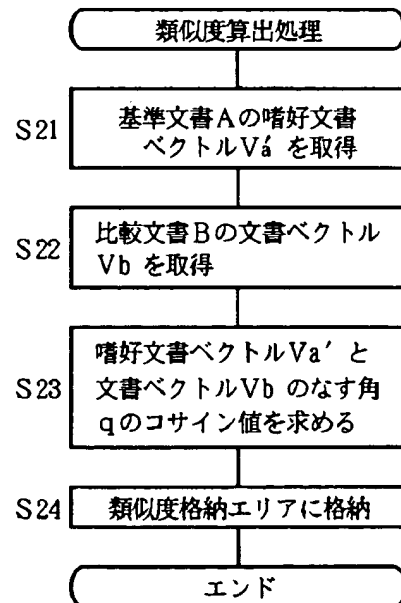
(74) 代理人 弁理士 川井 隆 (外1名)

(54) 【発明の名称】 文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法

(57) 【要約】

【課題】 ユーザーの嗜好を踏まえた文書処理を行うことのできる文書処理装置、文書処理プログラムを記憶した記憶媒体、及び文書処理方法を提供すること。

【解決手段】 過去の処理文書中の出現頻度等から処理重要語句(キーワード)を取得し、処理重要語句の処理文書中の出現頻度等からユーザー全体の嗜好を表すGPベクトルを取得する。また、類似度を取得する基準となる基準文書Aにおける処理重要語句の重要度を取得し、この重要度を要素とする文書ベクトルVaを取得し、GPベクトルと文書ベクトルVaの各要素を掛け合わせて嗜好文書ベクトルVa'を得る。この嗜好文書ベクトルVa'には、ユーザーの嗜好が反映されている。嗜好文書ベクトルVa'と他の文書Bの文書ベクトルVb、とでなす角qのコサイン値cos(q)を、基準文書Aと他の文書Bとの類似度とする。この類似度が大きいほど、他の文書はユーザーの嗜好に近いものとなる。



【特許請求の範囲】

【請求項1】 ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得手段と、

文書を取得する文書取得手段と、

前記文書取得手段により取得された文書の特徴付ける文書ベクトルを取得する文書ベクトル取得手段と、

前記文書ベクトル取得手段により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト手段とを具備することを特徴とする文書処理装置。

【請求項2】 前記嗜好ベクトル取得手段は、前記ユーザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成することを特徴とする請求項1に記載の文書処理装置。

【請求項3】 複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするGP行列を取得するGP行列取得手段を備え、

前記嗜好ベクトル取得手段は、前記GP行列により前記嗜好ベクトルを取得することを特徴とする請求項1または請求項2に記載の文書処理装置。

【請求項4】 前記GP行列取得手段は、ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書-キーワード行列と、ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書-ユーザー行列と、からGP行列を取得することを特徴とする請求項3に記載の文書処理装置。

【請求項5】 所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出手段を具備することを特徴とする請求項1から請求項4のうちのいずれか1の請求項に記載の文書処理装置。

【請求項6】 ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得機能と、

文書を取得する文書取得機能と、

前記文書取得機能により取得された文書の特徴付ける文書ベクトルを取得する文書ベクトル取得機能と、

前記文書ベクトル取得機能により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト機能をコンピュータに実現させるためのコンピュータ読み取り可能な文書処理プログラムが記憶された記憶媒体。

【請求項7】 前記嗜好ベクトル取得機能は、前記ユー

ザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成することを特徴とする請求項6に記載の文書処理プログラムが記憶された記憶媒体。

【請求項8】 複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするGP行列を取得するGP行列取得機能を備え、

前記嗜好ベクトル取得機能は、前記GP行列により前記嗜好ベクトルを取得することを特徴とする請求項6または請求項7に記載の文書処理プログラムが記憶された記憶媒体。

【請求項9】 前記GP行列取得機能は、ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書-キーワード行列と、

ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書-ユーザー行列と、からGP行列を取得することを特徴とする請求項8に記載の文書処理プログラムが記憶された記憶媒体。

【請求項10】 所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出機能を具備することを特徴とする請求項6から請求項9のうちのいずれか1の請求項に記載の文書処理プログラムが記憶された記憶媒体。

【請求項11】 ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルと、文書とを取得し、

前記文書の特徴付ける文書ベクトルを取得し、

前記文書ベクトルを前記嗜好ベクトルによりシフトさせることを特徴とする文書処理方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法に関し、更に詳細には、利用目的等のユーザーの嗜好を踏まえた類似文書の検索に関する。

【0002】

【従来の技術】従来の文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法による文書処理においては、複数の文書を検索・分類するために、それぞれの文書について当該文書の特徴付ける文書ベクトルを取得し、この文書ベクトルから文書間の類似度を算出することが行われている。

【0003】

【発明が解決しようとする課題】しかし、同一の文書でも、例えば営業用や技術資料用等の利用目的その他のユーザーの嗜好が異なると、重要部位等に差異が生じる。そして、上述のような文書ベクトルを用いた文書処理によって文書の検索や分類をしても、ユーザーの嗜好を踏まえた処理は行うことができないため、このような嗜好を加味した上での文書処理を行うことのできる文書処理装置や文書処理プログラムが記憶された記憶媒体、文書処理方法が望まれていた。

【0004】本発明は、上述のような課題を解決するためになされたもので、ユーザーの嗜好を踏まえた文書処理を行うことのできる文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法を提供することを目的とする。

【0005】

【課題を解決するための手段】請求項1に記載の発明は、図9に示すように、ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得手段101と、文書を取得する文書取得手段102と、前記文書取得手段102により取得された文書の特徴付けする文書ベクトルを取得する文書ベクトル取得手段103と、前記文書ベクトル取得手段により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト手段104とを具備する文書処理装置を提供することにより前記目的を達成するものである。請求項2に記載の発明は、図9に示すように、請求項1に記載の文書処理装置において、前記嗜好ベクトル取得手段101は、前記ユーザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成する文書処理装置を提供することにより前記目的を達成するものである。請求項3に記載の発明は、図10に示すように、請求項1または請求項2に記載の文書処理装置において、複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするGP行列を取得するGP行列取得手段105を備え、前記嗜好ベクトル取得手段101は、前記GP行列により前記嗜好ベクトルを取得する文書処理装置を提供することにより前記目的を達成するものである。請求項4に記載の発明は、図10に示すように、請求項3に記載の文書処理装置において、前記GP行列取得手段105は、ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書-キーワード行列と、ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書-ユーザー行列と、からGP行列を取得する文書処理装置を提供することにより前記目的を達成するものである。請求項5に記載の発明は、図11

に示すように、請求項1から請求項4のうちのいずれか1の請求項に記載の文書処理装置において、所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出手段106を具備する文書処理装置を提供することにより前記目的を達成するものである。請求項6に記載の発明は、図12に示すように、ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得機能201と、文書を取得する文書取得機能202と、前記文書取得機能202により取得された文書の特徴付けする文書ベクトルを取得する文書ベクトル取得機能203と、前記文書ベクトル取得機能203により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト機能204とをコンピュータに実現させるためのコンピュータ読み取り可能な文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成するものである。請求項7に記載の発明は、図12に示すように、請求項6に記載の記憶媒体において、前記嗜好ベクトル取得機能201は、前記ユーザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成するものである。請求項8に記載の発明は、図13に示すように、請求項6または請求項7に記載の記憶媒体において、複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするGP行列を取得するGP行列取得機能205を備え、前記嗜好ベクトル取得機能201は、前記GP行列により前記嗜好ベクトルを取得する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成する。請求項9に記載の発明は、図13に示すように、前記GP行列取得機能205は、ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書-キーワード行列と、ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書-ユーザー行列と、からGP行列を取得する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成するものである。請求項10に記載の発明は、図14に示すように、請求項6から請求項9のうちのいずれか1の請求項に記載の記憶媒体において、所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出機能206を具備する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を

達成するものである。請求項11に記載の発明は、図15に示すように、ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルと文書とを取得301し、前記文書の特徴付ける文書ベクトルを取得302し、前記文書ベクトルを前記嗜好ベクトルによりシフト303させる文書処理方法を提供することにより前記目的を達成するものである。

【0006】

【発明の実施の形態】以下、本発明の文書処理装置、文書処理プログラムが記憶された記憶媒体及び文書処理方法の好適な実施の形態について、図1から図8を参照して詳細に説明する。

(1) 実施形態の概要

本実施形態では、過去の処理文書中の出現頻度等から処理重要語句(キーワード) a, b, ...を取得し、処理重要語句の処理文書中の出現頻度、各処理文書の処理回数、処理したメンバーの重要度から、ユーザー全体の嗜好を表すGPベクトルを取得する。また、類似度取得する基準となる基準文書Aにおける処理重要語句 a, b, ...の重要度 $g(a)$, $g(b)$, ...を取得し、重要度 $g(a)$, $g(b)$, ...を要素とする文書ベクトル V_a を取得し、GPベクトルと文書ベクトル V_a , V_b , V_c , ...の各要素を掛け合わせて嗜好文書ベクトル V' aを得る。この嗜好文書ベクトルには、ユーザーの嗜好が反映されている。嗜好文書ベクトル V' aと他の文書の文書ベクトル V_b , とでなす角 q のコサイン値 $\cos(q)$ を、基準文書Aと他の文書Bとの類似度とする。この類似度が大きいほど、他の文書はユーザーの嗜好に近いものとなる。

【0007】(2) 実施形態の詳細

図1は、本発明の文書処理装置の一実施形態であり、本発明の文書処理プログラムが記憶された記憶媒体の一実施形態の該プログラムが読み取られたコンピュータの構成を表したブロック図である。この図1に示すように、文書処理装置(コンピュータ)は、装置全体を制御するための制御部11を備えている。この制御部11には、データバス等のバスライン21を介して、入力装置としてのキーボード12やマウス13、表示装置14、印刷装置15、記憶装置16、記憶媒体駆動装置17、通信制御装置18、入出力I/F19、及び文字認識装置20が接続されている。制御部11は、CPU111、ROM112、RAM113を備えている。ROM112は、CPU111が各種制御や演算を行うための各種プログラムやデータが予め格納されたリードオンリーメモリである。

【0008】RAM113は、CPU111にワーキングメモリとして使用されるランダムアクセスメモリである。このRAM113には、本実施形態による文書ベクトル取得処理を行うためのエリアとして、文書ベクトル取得の対象となる文書を格納する対象文書格納エリア1

131、キーワード格納エリア1132、文書ベクトル格納エリア1134が確保され、また、嗜好文書ベクトル取得処理を行うためのエリアとして、行列格納エリア1135、嗜好文書ベクトル格納エリア1136、類似度格納エリア1137その他の各種エリアが確保されるようになっている。

【0009】キーボード12は、かな文字を入力するためのかなキーやテンキー、各種機能を実行するための機能キー、カーソルキー、等の各種キーが配置されている。マウス13は、ポインティングデバイスであり、表示装置14に表示されたキーやアイコン等を左クリックすることで対応する機能の指定を行う入力装置である。表示装置14は、例えばCRTや液晶ディスプレイ等が使用される。この表示装置14には、文書ベクトルを取得する対象文書の内容や、本実施形態により算出された文書間の類似度、算出された類似度をもとに行った検索結果や分類結果等が表示されるようになっている。印刷装置15は、表示装置14に表示された文章や、記憶装置16の文書データベース164に格納された文書等の印刷を行うためのものである。この印刷装置としては、レーザプリンタ、ドットプリンタ、インクジェットプリンタ、ページプリンタ、感熱式プリンタ、熱転写式プリンタ、等の各種印刷装置が使用される。

【0010】記憶装置16は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータ等の各種情報を読み書きするための駆動装置で構成されている。この記憶装置16に使用される記憶媒体としては、主としてハードディスクが使用されるが、後述の記憶媒体駆動装置17で使用される各種記憶媒体のうちの読み書き可能な記憶媒体を使用するようにしてもよい。記憶装置16は、仮名漢字変換辞書161、プログラム格納部162、文書データベース164、重要語データベース165、行列データベース168、文書ベクトルデータベース166、嗜好文書ベクトルデータベース167、図示しないその他の格納部(例えば、この記憶装置16内に格納されているプログラムやデータ等をバックアップするための格納部)等を有している。プログラム格納部162には、本実施形態における嗜好文書ベクトル取得処理プログラム、類似度算出処理プログラム等の各種プログラムの他、仮名漢字変換辞書161を使用して入力された仮名文字列を漢字混り文に変換する仮名漢字変換プログラム等の各種プログラムが格納されている。

【0011】文書データベース164には、仮名漢字変換プログラムにより作成された文書や、他の装置で作成されて記憶媒体駆動装置17や通信制御装置18から読み込まれた文書が格納される。この文書データベース164に格納される各文書の形式は特に限定されるものではなく、テキスト形式の文書、HTML(Hyper Text Markup Language)形式の文書、JIS形式の文書等の各種形式の文書の格納が可能である。更にこの文書データ

ベース164には、文書処理したユーザー（処理者）及びその処理回数が各文書に対応付けて格納されている。前記処理回数は、所定期間毎に値を0にリセットされる。重要語データベース165には、前記所定期間内に処理した処理文書から抽出された重要語句（処理重要語句）及びその重要度（処理重要度）が格納される。

【0012】行列データベース168には、過去の所定期間に行われた文書処理の処理内容により取得される行列Ga、Gb、Gcが格納されている。文書ベクトルは、これらの行列Ga、Gb、Gcにより取得されるG

P（Group Personalize）行列をもとに、嗜好文書ベクトルに変換される。図2（a）～（c）は、行列Ga、Gb、Gcを示す説明図である。

【0013】行列Ga（文書－キーワード行列）は、図2（a）に示すように、前記所処理重要語句を行に、同処理文書を列にとった行列であり、各要素は処理重要語句の処理重要度f（x）を表している。行列Gb（文書－ユーザー行列）は、図2（b）に示すように、前記処理文書を行にとり、ユーザーのメンバーを列にとった行列であり、各要素は、メンバーが各文書を前記所定期間

内に処理した回数となっている。行列Gcは、図2（c）に示すように、行および列がともにユーザーのメンバーそれぞれの重要度係数を示している。行列Ga及び行列Gbは所定期間ごとに書き換えられ、行列Gcは操作者からの入力により適宜書き換えられる。

【0014】文書ベクトルデータベース166、及び嗜好文書ベクトルデータベース167には、本実施形態において類似度を算出する基準となる基準文書、及び、該基準文書に対する類似度を比較する対象文書それぞれの文書ベクトル及び嗜好文書ベクトルが格納される。

【0015】記憶媒体駆動装置17は、CPU111が外部の記憶媒体からコンピュータプログラムや文書を含むデータ等を読み込むための駆動装置である。記憶媒体に記憶されているコンピュータプログラムには、本実施形態の文書処理装置により実行される各種処理のためのプログラム、および、そこで使用される辞書、データ等も含まれる。ここで、記憶媒体とは、コンピュータプログラムやデータ等が記憶される記憶媒体をいい、具体的には、フロッピーディスク、ハードディスク、磁気テープ等の磁気記憶媒体、メモリチップやICカード等の半導体記憶媒体、CD-ROMやMO、PD（相変化書換型光ディスク）等の光学的に情報が読み取られる記憶媒体、紙カードや紙テープ等の用紙（および、用紙に相当する機能を持った媒体）を用いた記憶媒体、その他各種方法でコンピュータプログラム等が記憶される記憶媒体が含まれる。本実施形態の文書処理装置において使用される記憶媒体としては、主として、CD-ROMやフロッピーディスクが使用される。記憶媒体駆動装置17は、これらの各種記憶媒体からコンピュータプログラムを読み込む他に、フロッピーディスクのような書き込み

可能な記憶媒体に対してRAM113や記憶装置16に格納されているデータ等を書き込むことが可能である。

【0016】本実施形態の文書処理装置では、制御部11のCPU111が、記憶媒体駆動装置17にセットされた外部の記憶媒体からコンピュータプログラムを読み込んで、記憶装置16の各部に格納（インストール）する。そして、本実施形態による類似度算出等の各種処理を実行する場合、記憶装置16から該当プログラムをRAM113に読み込み、実行するようになっている。但し、記憶装置16からではなく、記憶媒体駆動装置17により外部の記憶媒体から直接RAM113に読み込んで実行することも可能である。また、文書処理装置によっては、本実施形態の自動要約処理プログラム等を予めROM112に記憶しておき、これをCPU111が実行するようにしてもよい。

【0017】通信制御装置18は、他のパーソナルコンピュータやワードプロセッサ等との間でテキスト形式やHTML形式等の各種形式の文書やビットマップデータ等の各種データの送受信を行うことができるようになっている。入出力I/F19は、音声や音楽等の出力を行うスピーカ等の各種機器を接続するためのインターフェースである。文字認識装置20は、用紙等に記載された文字をテキスト形式やHTML等の各種形式で認識する装置であり、イメージスキャナや文字認識プログラム等で構成されている。

【0018】本実施形態では、キーボード12の入力操作により作成した文書（RAM113の所定格納エリアに格納）の他、外部で作成して所定の記憶媒体に格納した文書で記憶媒体駆動装置17から読み込んだ文書、予め文書データベース164に格納されている文書、通信制御装置18からダウンロードした文書、及び文字認識装置20で文字認識した文書、等の各種文書を対象文書として取得することが可能である。

【0019】次に、上述のような構成の文書処理装置による嗜好文書ベクトル取得処理及び類似度算出処理について図面を参照して説明する。

【0020】本実施形態においては、所定期間毎に、該所定期間内に行われた文書処理の処理内容に基づいて新たな処理重要語句及び処理重要度が取得され、行列データベース168内の行列Ga及び行列Gbが書き換えられる。

【0021】図3は、行列Ga、Gb書き換え処理の動作を表したフローチャートである。CPU111は、所定期間内に処理された文書（処理文書）を文書データベース164から順次取得してRAM113の所定作業領域に格納し（ステップ11）、各処理文書についての重要語句（処理重要語句）及びその重要度（処理重要度）を取得する（ステップ12）。

【0022】図4は処理重要語句・処理重要度取得処理の動作を表したフローチャートである。図4に示すよう

に、CPU111は、文書データベース164から取得した処理文書について、各処理文書毎に形態素解析を行うことで自立語を抽出する(ステップ121)と共に、名詞句、複合名詞句等を含めた候補語(句)を処理文書から抽出する(ステップ122)。次に、抽出した候補語(句)の処理文書での出現頻度、評価関数から、各候補語(句)の処理重要度 $f(x)$ を取得する(ステップ123)。ここで、評価関数としては、例えば、所定の重要語句が予め指定されている場合にはその重要語句に対する重み付け、単語、名詞句、複合名詞句等の候補語(句)の種類による重み付け等が使用される。

【0023】さらにCPU111は、取得した処理重要度 $f(x)$ の値をもとに候補語(句)から処理重要語句 a, b, \dots を取得し(ステップ124)、この処理重要語句 a, b, \dots 及びその処理重要度 $f(a), f(b), \dots$ を重要語データベース165に格納する(ステップ125)。

すべての処理文書について、処理重要語句及びその処理重要度を取得すると、図4に示す行列 G_a, G_b 書き換え処理ルーチンへリターンする。

【0024】続いて、CPU111は、行列データベース168の行列 G_a を、前記処理重要語句 a, b, \dots を行に、前記所定期間の処理文書を列に、また処理重要度 $f(a), f(b), \dots$ を各要素にとったものを書き換える(ステップ13)。このとき、行列 G_a の行数は、各処理文書の処理重要語句の和集合の数とし、各処理文書において含まれていない処理重要語句については、その処理重要度 $f(x)$ は0と定義される。

【0025】例えば図2において、処理文書Bの処理重要語は「重要、重要語、重要度、…」、処理文書Cの処理重要語句は「重要、…、政治、…」であり、これらの処理重要語句に対応する処理重要度は、処理文書Bについては(1, 18, 19, …)、処理文書Cについては(18, …, 21, …)である。これに対して行列 G_a においては、その行は「重要、重要語、重要度、…、政治、…」とし、両文書の列における要素値はつぎの通り定義される。

処理文書Bの列=(1, 18, 19, …, 0, …)

処理文書Cの列=(18, 0, 0, …, 21, …)

【0026】また、CPU111は、文書データベース164から、各処理文書の処理回数を取得し(ステップ14)、行列 G_b を、所定期間内の処理文書を行に、文書データベース164から取得した処理回数を各要素としたものを書き換えて(ステップ15)、行列 G_a, G_b 書き換え処理を終了する。

【0027】図5は、本実施形態による文書嗜好ベクトル取得処理の動作を示すフローチャートである。嗜好文書ベクトル取得に際しては、嗜好文書ベクトル取得の対象となる文書(対象文書)を取得し、RAM113の対象文書格納エリア1131に格納する(ステップ2

1)。対象文書は、ユーザの指示に従ってRAM113、記憶装置16の文書データベース164、記憶媒体駆動装置17、または通信制御装置18(パソコン通信、インターネット等の通信による場合)から取得する。

【0028】次にCPU111は、対象文書中から行列 G_a の処理重要語句を抽出する(ステップ22)。次に、抽出した処理重要語句の対象文書中での出現頻度、評価関数等から、重要度 $g(y)$ を取得する(ステップ23)。ここで、評価関数としては、例えば、処理重要語句に対する予め指定されている重み付け、単語、名詞句、複合名詞句等の候補語(句)の種類による重み付け等が使用される。

【0029】そして、この処理重要語句 a, b, \dots の対象文書における重要度 $g(a), g(b), \dots$ を要素とする文書ベクトル V を取得する(ステップ24)。

【0030】文書ベクトル V を取得すると、CPU111は、行列データベース168から行列 G_a, G_b, G_c を取得し、次の式に従って、 GP 行列を求める(ステップ25)。

$$GP = G_a \cdot G_b \cdot G_c$$

従って、本実施形態における GP 行列は、 G_a 行列の次元合わせを行った行をそのまま行にとり、ユーザーの各メンバーを列にとってなっており、 GP 行列の各要素は、メンバー毎の過去の文書処理における処理重要語句の処理重要度 $f(x)$ に各メンバーの重要度を加味して表した数値となっている。

【0031】 GP 行列が取得されると、続いてCPU111は、この GP 行列をもとに GP ベクトル(嗜好ベクトル)を取得する(ステップ26)。図6は、 GP 行列から GP ベクトルを算出する行程を概念的に説明する説明図である。

【0032】CPU111は、まず、 GP 行列の各要素 $g_{ij}(i=1 \sim \text{メンバー数}m, j=1 \sim \text{処理重要語句の和集合の数}k)$ の各行毎の要素の平均値を算出して列ベクトル(総 GP ベクトル)を得る(図6(1)→

(2))。この総 GP ベクトルは、各要素 g_i が処理重要語句毎のユーザーグループ全体における過去の文書処理での出現頻度(但し各処理重要語句の予め決められた処理重要語句の重み等や、メンバーの重要度が加味されている)を反映した数値となっている。CPU111は、更に、この総 GP ベクトルの各要素 g_i を文書の処理回数の総数で割って、1列の GP ベクトルを得る(図6(2)→(3))。この様に、総 GP ベクトルを文書の処理回数の総数で割るのは、行列 G_b に文書の処理回数が要素として含まれており、処理回数が増えるに従って GP ベクトルが大きくなっていくのを回避するためである。

【0033】そして、CPU111は、 GP ベクトルの各要素とこの各要素に対応する文書ベクトル V の要素と

を掛け合わせて、嗜好文書ベクトル V' を得る。嗜好文書ベクトル V' は、嗜好文書ベクトルデータベース167に格納して(ステップ26)、嗜好文書ベクトル取得処理を終了する。

【0034】図7は、文書ベクトルを嗜好文書ベクトルにシフトさせた状態を概念的に示す説明図である。尚、この説明図では、表示の都合上処理重要語句をX、Yの2つとして説明するが、処理重要語句の数が異なっている、文書ベクトルを嗜好文書ベクトルにシフトさせた状態については本質的に同様である。図7に示すように、文書ベクトル $V_p = (0, 1)$ 、文書ベクトル $V_q = (2, 1)$ 、及び文書ベクトル $V_r = (8, 1)$ をGPベクトル $= (1, 2)$ によりシフトさせたものである。文書ベクトル V_p 、文書ベクトル V_q 、文書ベクトル V_r は、GPベクトルにより、それぞれ嗜好文書ベクトル $V_p' = (0, 2)$ 、 $V_q' = (2, 2)$ 、 $V_r = (8, 2)$ にシフトされる。このように、出現頻度等によって決定される要素値により、文書ベクトルから嗜好文書ベクトルへ様々な角度でシフトされる。

【0035】次に、本実施形態による類似度算出処理について図8を参照して説明する。図8は、類似度算出処理の動作を示すフローチャートである。

【0036】類似度算出に際しては、CPU111は、類似度を算出する基準となる文書(基準文書A)についての文書嗜好ベクトル $V'a$ を取得する(ステップ21)。基準文書Aについての文書嗜好ベクトル $V'a$ は、上述の動作によって、または既に嗜好文書ベクトルデータベース167に格納されている場合にはこの文書嗜好ベクトルデータベース167から読み出して、取得する。尚、基準文書Aは、文書データベース164に格納されている文書等から処理時に選択しても、予め決定されているものを用いてもよい。また、基準文書Aとしては、1つの文書の他、複数の文書よりなる文書群や、文書群からクラスタリング処理により自動抽出した文書群を用いることもできる。

【0037】また、基準文書Aとの類似度を算出する比較文書Bについての文書ベクトル Vb を取得する(ステップ22)。

【0038】次に、CPU111は、基準文書Aと比較文書Bとの類似度 Sb を、基準文書Aの嗜好文書ベクトル Va' と比較文書Bの文書ベクトル Vb 間の角度 q に依存するコサインにより求める(ステップ23)。すなわち、比較文書Bの基準文書Aに対する類似度 Sb は、嗜好文書ベクトル Va' と文書ベクトル Vb の間の角度を q 、嗜好文書ベクトル Va' と文書ベクトル Vb の内積を $Va' \cdot Vb$ 、嗜好文書ベクトル Va' 、文書ベクトル Vb の大きさをそれぞれ $|Va'|$ 、 $|Vb|$ とした場合、次の数式1により求まる。

【0039】

【数式1】類似度 $Sb = \cos(q) = (Va' \cdot V$

$b) / (|Va'| \times |Vb|)$

【0040】この類似度 Sb の値は $-1 \leq Sb \leq 1$ までの値をとり、1に近いほど比較文書Bの文書ベクトル Vb と基準文書Aの嗜好文書ベクトル Va' との向きが近く、文書ベクトル Vb が嗜好文書ベクトル Va' に類似し、比較文書Bが、ユーザーの嗜好に近いと考えることができる。

【0041】CPU111は、求められた類似度 Sb を比較文書Bと対応させて類似度格納エリア1137に格納して(ステップ24)、類似度算出処理を終了する。

【0042】この様に、本実施形態では、ユーザーの処理文書中における処理重要語句の頻出頻度をもとに、基準文書Aの文書ベクトル V をシフト(文書ベクトルの各要素値を変換)してユーザーの嗜好を加味した嗜好文書ベクトル V' を取得し、この嗜好文書ベクトル V' に対する他の文書の類似度が算出される。従って、この類似度は、ユーザーの興味や注目度、目的等の嗜好ユーザーの嗜好に対する指標となるので、この類似度に基づいて文書の分類や検索を行うことにより、ユーザーの嗜好を反映した分類や検索が可能となる。また、ユーザーの興味や注目度等に合う文書を選択して配信することが可能となる。本実施形態によると、GP行列を用いた変換によって文書ベクトル V を嗜好文書ベクトル V' にシフトさせているので、計算処理が簡単であり、ベクトル空間法を採用したコア・エンジンを備えた一般の文書処理装置に容易に適用することが可能である。

【0043】本実施形態によると、文書ベクトルを嗜好文書ベクトルにシフトさせるGPベクトル(嗜好ベクトル)のもととなるGP行列を、表現すべき特徴毎の単純な観点で構成した行列 Ga 、 Gb 、 Gc の掛け合わせて求めているので、様々な特徴を考慮に入れたGP行列を容易に構成して文書ベクトル V をシフトさせることが可能である。本実施形態によると、文書ベクトルを嗜好文書ベクトルにシフトさせるGPベクトルのもととなるGP行列は、各列がユーザーのメンバーの興味を反映しているので、ユーザーを数グループに分割した該グループのGP行列や個々のメンバーのGP行列(ベクトル)を容易に得ることができる。本実施形態によると、GP行列がユーザーの過去に処理した文書をもとに適宜書き換えられている行列 Ga 、 Gb 、 Gc をもとに取得されているので、文書ベクトル V がユーザーの嗜好の経時的揭示変化に対応した嗜好文書ベクトル V' にシフトされ、ユーザーの嗜好の変遷に追従した類似度の算出および検索・分類等の処理が可能となる。

【0044】尚、本発明は、上述の実施形態に限定されるものではなく、本発明の趣旨を逸脱しない限りにおいて適宜変更が可能である。上述の実施形態においては文書処理装置としてコンピュータを用いているが、コンピュータに限定されるものではなく、ワードプロセッサ等であってもよい。上述の実施形態においてはGP行列

は、メンバー毎の過去の文書処理回数（行列G a）と各文書における処理重要語句の出現頻度（行列G b）、および各メンバーの重要度（行列G c）とから取得されているが、メンバー毎の過去の文書処理回数（行列G a）と各文書における処理重要語句の出現頻度（行列G b）のみにより取得されてもよい。また、例えば、各文書の処理時間や、他の文書作成に引用された件数、リンク付けされている数等も加味して取得されてもよい。更に、GP行列を上述の実施形態と同様に行列G a～行列G c等の行列の掛け合わせから取得する場合において、行列G a～行列G c等の各行列の要素はそれぞれ処理重要語句の文書中の出現頻度や、メンバーが各文書を処理した回数を反映した数値となっていればよく、直接出現頻度や処理回数そのものを表していなくてもよい。

【0045】上述の実施形態においては行列G a～G cは過去の文書処理内容から取得されているが、ユーザーが取得して行列データベース168に直接入力してもよい。上述の実施形態においては行列G a～G cは所定期間毎に書き換えられているが、文書処理を行う毎に、または所定回数の文書処理を行う毎等にも書き換えてもよい。GPベクトルを表示装置に表示するGPベクトル表示手段を備え、ユーザーやユーザーメンバーの嗜好を視覚的に把握できるようにしてもよい。この場合、GPベクトルを行列データベースまたは専用のGPベクトルデータベースに経時順に格納しておき、経時変化も把握できるようにしてもよい。

【0046】説明した実施形態では処理重要語句や処理重要度を取得する手法として図4のフローチャートに従った方法を1例にして説明したが、本発明でこの方法に限られるものではなく、文書中から処理重要語句を抽出する方法や、処理重要度の決定方法等については、公知の各種方法により置き換えることが可能である。更に、2つの文書嗜好ベクトルの類似度の算出方法については、数式1により類似度を算出することとしたが、この数式に限定されるものではなく、文書嗜好ベクトル相互間の類似関係を表すことが可能であれば他の数式により類似度を算出することも可能である。算出した類似度の表示は、類似度の操作者からの入力により類似閾値を取得し、当該類似閾値よりも高い類似度を備えた対象文書のみを表示させたり、類似度の高いうちから10個の文書のみを表示させたりすることもできる。また、類似度の高い順ではなく、あいうえお順等に表示された対象文書名とともに表示してもよい。更に、類似度表示は、操作者からの命令のあったときのみに表示させるようにしたり、表示装置には表示させずに印刷させることとしてもよい。

【0047】説明した実施形態は日本語で作成された文書に限られるものではなく、あらゆる言語で作成された文書を対象とすることが可能である。その場合、対象となる文書が作成された言語用の形態素解析アルゴリズム

等を使用するといった、本発明の構成には影響のない部分を変更するだけでよい。

【0048】なお、以上の実施形態において説明した、各装置、各部、各動作、各処理等に対しては、それらを含む上位概念としての各手段（～手段）により、実施形態を構成することが可能である。例えば、「文書データベース164から、各処理文書の処理回数を取得し（ステップ14）」との記載に対して文書の処理回数を記憶する処理回数データベースを文書データベース164とは別途に構成したり、「処理回数取得手段」を構成したり、「抽出した候補語（句）の処理文書での出現頻度、評価関数から、各候補語（句）の処理重要度 $f(x)$ を取得する（ステップ123）」との記載に対して、「処理重要語句取得手段」を構成するようにしてもよい。同様に、その他各種動作に対して「～（動作）手段」等の上位概念で実施形態を構成するようにしてもよい。

【0049】

【発明の効果】以上説明したように、本発明によれば、嗜好文書ベクトル取得手段により文書ベクトルをユーザーの嗜好を加味した嗜好文書ベクトルにシフトさせ、この嗜好文書ベクトルに対する類似度を取得することにより、ユーザーの興味や注目度、目的等の嗜好に対する文書の類似度が取得でき、この類似度に基づいて分類や検索を行うことにより、ユーザーの興味や注目度、目的等の嗜好を反映した分類や検索、配信等の文書処理が可能となる。

【図面の簡単な説明】

【図1】本発明の文書処理装置の一実施形態であり、本発明の文書処理プログラムが記憶された記憶媒体の一実施形態の該プログラムが読み取られたコンピュータの構成を表したブロック図である。

【図2】図1の実施形態における行列G a、G b、G cを示す説明図である。

【図3】図1の実施形態による行列G a、G b書き換え処理の動作を表したフローチャートである。

【図4】図1の実施形態による処理重要語句・処理重要度取得処理の動作を表したフローチャートである。

【図5】図1の実施形態による嗜好文書ベクトル取得処理の動作を示すフローチャートである。

【図6】図1の実施形態におけるGP行列からGPベクトルを算出する行程を概念的に説明する説明図である。

【図7】図1の実施形態における文書ベクトルを嗜好文書ベクトルにシフトさせた状態を概念的に説明する説明図である。

【図8】図1の実施形態による類似度算出処理の動作を示すフローチャートである。

【図9】請求項1に記載した発明のクレーム対応図である。

【図10】請求項3に記載した発明のクレーム対応図である。

【図11】請求項5に記載した発明のクレーム対応図である。

【図12】請求項6に記載した発明のクレーム対応図である。

【図13】請求項8に記載した発明のクレーム対応図である。

【図14】請求項10に記載した発明のクレーム対応図である。

【図15】請求項11に記載した発明のクレーム対応図である。

【符号の説明】

11	制御部
112	ROM
113	RAM
1131	対象文書格納エリア
1132	キーワード格納エリア
1134	文書ベクトル格納エリア
1135	行列格納エリア
1136	嗜好文書ベクトル格納エリア
1137	類似度格納エリア
12	キーボード
13	マウス
14	表示装置
15	印刷装置

*16	記憶装置
161	仮名漢字変換辞書
162	プログラム格納部
164	文書データベース
165	重要語データベース
166	文書ベクトルデータベース
167	嗜好文書ベクトルデータベース
168	行列データベース
17	記憶媒体駆動装置
18	通信制御装置
19	入出力I/F
101	嗜好ベクトル取得手段
102	文書取得手段
103	文書ベクトル取得手段
104	シフト手段
105	GP行列取得手段
106	類似度算出手段
201	嗜好ベクトル取得機能
202	文書取得機能
203	文書ベクトル取得機能
204	シフト機能
205	GP行列取得機能
206	類似度算出機能

*

【図2】

【図3】

【図4】

【図5】

(a) 行列 Ga (キーワード、文書)

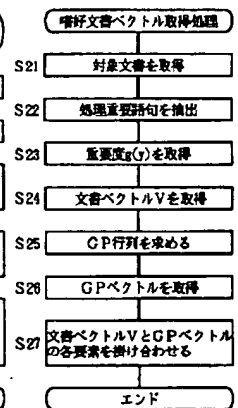
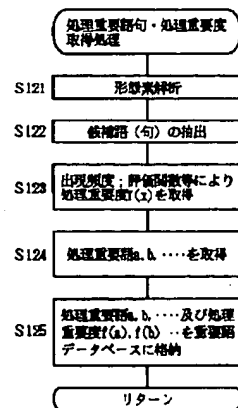
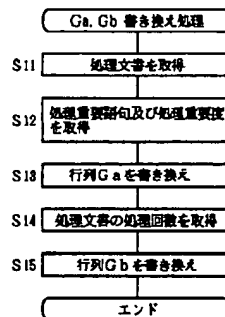
	文書A	文書B	文書C...
重要	2	1	18 ...
重要語	5	18	0 ...
重要度	1	19	0 ...
政治	2	0	21 ...
...

(b) 行列 Gb (文書、処理者)

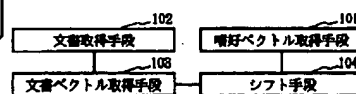
	星太郎	花園美子	黒井三四郎	意見五郎
文書A	1	0	1	0
文書B	1	1	2	0
文書C	1	1	1	1
...

(c) 行列 Gc (処理者の重要度)

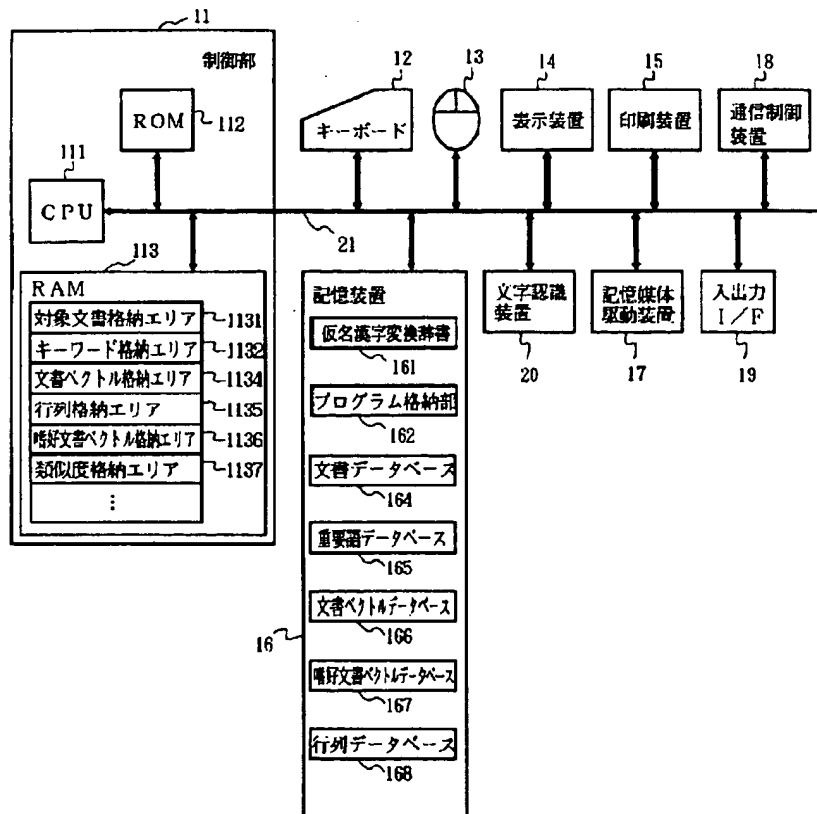
	星太郎	花園美子	黒井三四郎	意見五郎
星太郎	1.5	0	0	0
花園美子	0	0.8	0	0
黒井三四郎	0	0	1.3	0
意見五郎	0	0	0	1.1



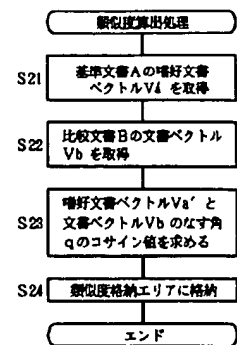
【図9】



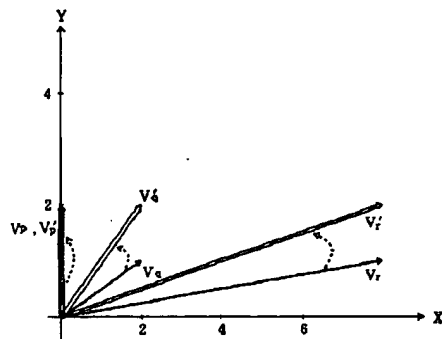
【図1】



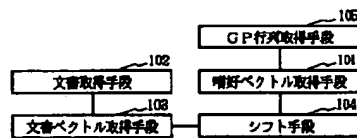
【図8】



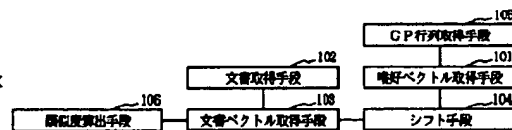
【図7】



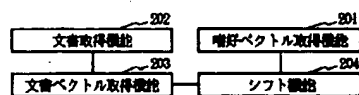
【図10】



【図11】



【図12】



【図6】

$$\text{キーワード} \begin{pmatrix} g_{11} & g_{12} & g_{13} & \cdots & g_{1n} \\ g_{21} & g_{22} & g_{23} & \cdots & g_{2n} \\ g_{31} & g_{32} & g_{33} & \cdots & g_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{k1} & g_{k2} & g_{k3} & \cdots & g_{kn} \end{pmatrix} \quad (1)$$

GP行列

平均化

$$\begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \\ g_k \end{pmatrix} \quad (2) \quad \left(g_i = \frac{g_{i1} + g_{i2} + \cdots + g_{in}}{n} ; i = 1 \sim k \right)$$

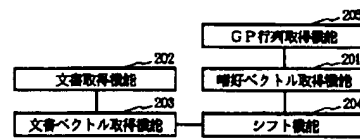
縮GPベクトル

規格化

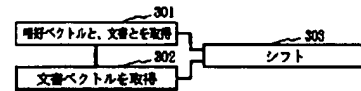
$$\begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_k \end{pmatrix} \quad (3) \quad \left(h_i = \frac{g_i}{\text{各文書の総出現数の合計}} \right)$$

GPベクトル

【図13】



【図15】



【図14】

